New Methods for Representing and Interacting with Qualitative Geographic Information

Contract #: W912HZ-12-P-0334

Contract Period: Sept. 30, 2012 – June 30, 2013

Principal Investigators:

Dr. Alan M. MacEachren, GeoVISTA Center, Penn State University

Dr. Prasenjit Mitra, IST & GeoVISTA Center, Penn State University

Dr. Anthony Robinson, GeoVISTA Center, Penn State University

Report on New Methods for Representing and Interacting with Qualitative Geographic Information, Stage 2: Task Group 1 – Core Re-engineering and Place-based Use Case

Alan M. MacEachren, Alexander Savelyev, Scott Pezanowski, Anthony C. Robinson, and Prasenjit Mitra

< maceachren, savelyev, spezanowski, arobinson >@psu.edu; pmitra@ist.psu.edu

GeoVISTA Center, Department of Geography, The Pennsylvania State University

Submitted, June, 30, 2013

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* **06-14-2013**	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* Sept. 30, 2012 – June 30, 2013

4. TITLE AND SUBTITLE

Report on New Methods for Representing and Interacting with Qualitative Geographic Information, Stage 2: ***Task Group 1 – Core Re-engineering and Place-based Use Case***

5a. CONTRACT NUMBER:
W912HZ-12-P-0334

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Alan M. MacEachren, Alexander Savelyev, Scott Pezanowski, Anthony C. Robinson, and Prasenjit Mitra

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

PENNSYLVANIA STATE UNIVERSITY , THE
408 OLD MAIN
UNIVERSITY PARK PA 16802-1505

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Engineer Research and Development Center (ERDC)
Topographic Engineering Center (TEC)
7701 Telegraph Road
Alexandria, VA 22135-3864

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; Distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report documents Pennsylvania State University's (PSU) research on place-focused analysis of microblogs, specifically Twitter. The first section of the report identifies three categories of location information related to tweets: the location where a tweet originates, places mentioned in tweets, and the profile location of the tweeter. The report summarizes previous research on location information and contributes new insights resulting from additional PSU analysis. The second section of the report focuses on the visualization of place information. Place trees, place clouds, and place-coreferencing are introduced in the context of the SensePlace2 system. The third section of the report summarizes the SensePlace2 system architecture enhancements required to support the dynamic visualization of Twitter location information, including the development of an innovative user interface coordination mechanism and implementation of faceted search using Apache Solr.

15. SUBJECT TERMS

geovisualization, visual analytics, social media, microblogs, cartography, qualitative geographic information, text analytics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
SAR	SAR	SAR	SAR	35	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Abstract

This report documents Pennsylvania State University's (PSU) research on place-focused analysis of microblogs, specifically Twitter. The first section of the report identifies three categories of location information related to tweets: the location where a tweet originates, places mentioned in tweets, and the profile location of the tweeter. The report summarizes previous research on location information and contributes new insights resulting from additional PSU analysis. The second section of the report focuses on the visualization of place information. Place trees, place clouds, and place-coreferencing are introduced in the context of the SensePlace2 system. The third section of the report summarizes the SensePlace2 system architecture enhancements required to support the dynamic visualization of Twitter location information, including the development of an innovative user interface coordination mechanism and implementation of faceted search using Apache Solr.

1 Introduction

Microblogs (blogs with very brief entries, usually with a fixed character limit) are used by a wide cross-section of the population, in countries around the world, to post diverse kinds of information using a limited number of characters per post. A large proportion of posts are probably of interest only to the creator and their immediate friends (or perhaps only to the creator). But with more than 400 million posts per day on Twitter alone (Tsukayama, 2013), even the small proportion that are of interest beyond the creator represents a very large resource. For Twitter, prior estimates (e.g., Leetaru et al., 2013) and our own past analyses of collected tweets (MacEachren et al., 2011a) suggest that between 1% and 2% of posts include a geographic location that reveals where the tweet was submitted from. Twitter posts also frequently mention locations in the messages themselves and our prior work has demonstrated that these can be leveraged to support situational awareness (MacEachren et al., 2011a). The most recent estimate using a very large sample (1.5 Billion tweets) found that 3% have geolocation when the "geo" metadata field and the "user-defined location" fields are combined (Leetaru et al., 2013). In addition to the fact that most past analysis seems to have ignored the user-defined location field, there is evidence that during crisis events, the percentage of relevant posts that include geolocation increases ((e.g., for earthquake posts, 15% has been observed, Guy et al., 2010)). Thus, Twitter and other microblog platforms (e.g., Sina Weibo in China) represent a rich source of potential open-source geographic information to complement more traditional information sources. In particular, it is a source that may act to fill in many local gaps where there is no data collected by traditional sources. Also, the value of Twitter is in its real-time nature. When events are happening, people go to Twitter to follow them in real-time and to post what they know.

In this stage of research, we have focused specifically on supporting place-based analysis that leverages Twitter posts as an open data source. The overall objectives are to: understand the characteristics of location references available with Twitter (that include references to place in the tweet, locations that the tweet is from if users opt-in to the location feature, indication of locations in the profile location field and in other metadata such as the time zone field); develop methods to utilize the various kinds of location separately and together; and demonstrate the potential utility of Twitter for place-based analysis through implementation of the methods within the SensePlace2 web application under development in the GeoVISTA Center at Penn State.

This report contains four sections that: introduce types of place reference found in Twitter and summarize preliminary findings about the types; outline a range of query and visualization methods developed and implemented to support place-centric analysis with Twitter data; describe the system architecture implemented to enable dynamic control, query, and multiview visualization; and present some conclusions and ideas for future research.

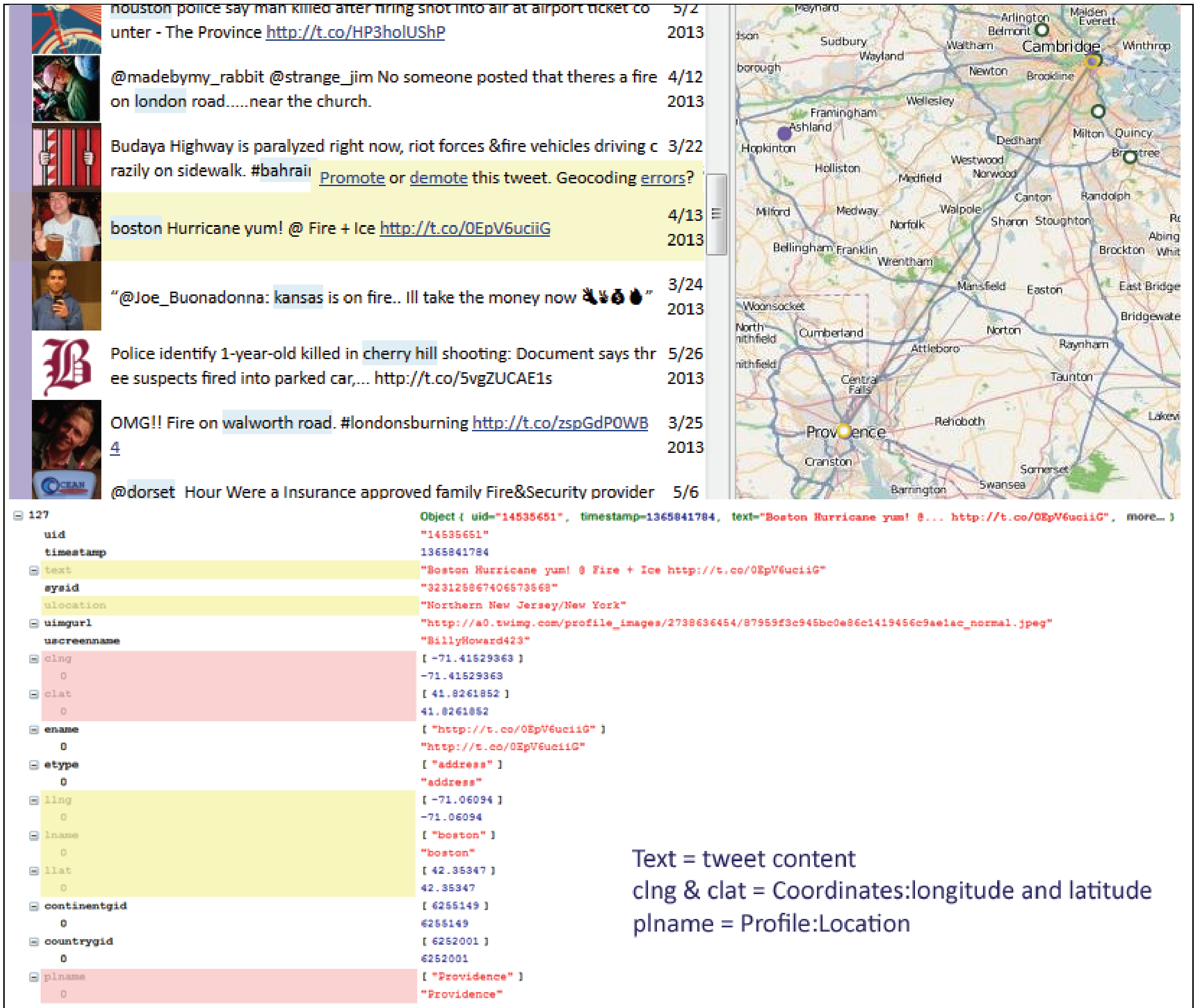
2 Types of Places in Twitter

We identify three primary categories of location information related to tweets: (a) the place *from* which a post is broadcast, (b) the place to which the tweeter claims *affiliation*, and (c) the place or places that the post is *about* (Figure 1). In addition to potential variation in accuracy, each can vary in geographic precision (with reference to coordinates, street intersections, or buildings at one extreme and reference to countries or continents at the other). Each is also represented by a range from spatially ambiguous to unambiguous specifications (e.g., “downtown” versus “Centre County”).

Below, we review related research on acquiring and using references to the *from*, *affiliated*, and *about* locations with which tweets are associated and then present key insights we have concerning geolocation in Twitter based on multiple Twitter data collection and processing activities. First, we provide more detail on each type of location information that is potentially available.

The place a tweet is from:

For the small percentage of tweets that come with geolocation, that location can be specified precisely with coordinates (either generated by GPS-enabled devices used to send the tweet or from the IP address of the computer used to post the tweet) or with a place name specified manually by the user (this location specification option is only available when tweets are posted from a web browser; users can insert a neighborhood, city, state, or country but there is nothing to prevent them from inserting a non-place).



houston police say man killed after firing shot into air at airport ticket counter - The Province http://t.co/HP3holUShP 5/2 2013

@madebymy_rabbit @strange_jim No someone posted that theres a fire on london road.....near the church. 4/12 2013

Budaya Highway is paralyzed right now, riot forces &fire vehicles driving crazily on sidewalk. #bahrai Promote or demote this tweet. Geocoding errors? 3/22

boston Hurricane yum! @ Fire + Ice http://t.co/0EpV6uciiG 4/13 2013

"@Joe_Buonadonna: kansas is on fire.. Ill take the money now 🔪💸💰💣" 3/24 2013

Police identify 1-year-old killed in cherry hill shooting: Document says three suspects fired into parked car,... http://t.co/5vgZUCAE1s 5/26 2013

OMG!! Fire on walworth road. #londonsburning http://t.co/zspGdP0WB4 3/25 2013

@dorset Hour Were a Insurance approved family Fire&Security provider 5/6

```
⊟ 127                 Object { uid="14535651", timestamp=1365841784, text="Boston Hurricane yum! @... http://t.co/0EpV6uciiG", more... }
    uid               "14535651"
    timestamp         1365841784
  ⊟ text              "Boston Hurricane yum! @ Fire + Ice http://t.co/0EpV6uciiG"
    sysid             "323125867406573568"
    ulocation         "Northern New Jersey/New York"
  ⊟ uimgurl           "http://a0.twimg.com/profile_images/2738636454/87959f3c945bc0e86c1419456c9ae1ac_normal.jpeg"
    uscreenname       "BillyHoward423"
  ⊟ clng              [ -71.41529363 ]
      0               -71.41529363
  ⊟ clat              [ 41.8261852 ]
      0               41.8261852
  ⊟ ename             [ "http://t.co/0EpV6uciiG" ]
      0               "http://t.co/0EpV6uciiG"
  ⊟ etype             [ "address" ]
      0               "address"
  ⊟ llng              [ -71.06094 ]
      0               -71.06094
  ⊟ lname             [ "boston" ]
      0               "boston"
  ⊟ llat              [ 42.35347 ]
      0               42.35347
  ⊟ continentgid      [ 6255149 ]
      0               6255149
  ⊟ countrygid        [ 6252001 ]
      0               6252001
  ⊟ plname            [ "Providence" ]
      0               "Providence"
```

Text = tweet content
clng & clat = Coordinates:longitude and latitude
plname = Profile:Location

Figure 1. A tweet highlighted (above) and key components of tweet metada retrieved by the Twitter API, including the tweet text ("text"), a location specified in the user profile (plname), and the location that the tweet was posted from (clng & clat).

Tweet author's place affiliation:

Users can claim affiliation with a location by inserting that location in their profile and/or by selecting the time-zone they prefer to be affiliated with (which, of course, is less precise). Not all users include a location in their profile or pick a time zone and many users who specify a location do not provide one that is meaningful (see discussion below). Also, since it is unlikely that many users continually update their profile when they move around, the place or time zone in the profile is best considered as the location to which the user claims affiliation (this association is a hypothesis, since the extent to which named profile locations match times zones has not been investigated); city-level profile location, however, may be a good guess on where a post is from (depending on how mobile the user is).

Place(s) a tweet is about:

Where a tweet is about is not usually as explicit as where it is from. Estimating the *about* location for a tweet is comparable to efforts to determine the geographic footprint of larger documents (Jones and Purves, 2008), except that the challenge of assembling context to make the estimate is much harder due to the limited information in the tweet itself. The locations to which a tweet refers typically need to be inferred based upon information contained in the body of the tweet (e.g., direct use of place names), the context derived from tweet metadata, or through the identification of relationships between the tweet and other information (e.g., that on May 20 or 21, 2013, mentions of “destroyed” had a high probability of being associated with Moore, OK, particularly in a tweet replying to another that mentioned Moore, OK). As outlined below, determining where tweets are about is an active area of research that addresses multiple aspects of the challenge from a range of perspectives.

2.1 Related research

As noted above, we have identified three categories of location associated with tweets: (a) the place the tweet is *from*, (b) the tweet author’s place *affiliation*, and (c) the place(s) the tweet is *about*. Ikawa, et al (Ikawa et al., 2013) present a similar typology that includes: (a) user’s current location, (b) user’s profile location, (c) locations in text, and (d) focused locations. The first two match our categories. The third is somewhat narrower, only including locations mentioned by name in the text of the tweet. Our *about* category includes the potential to infer locations that are not mentioned explicitly. Ikawa and colleagues’ fourth category is hard to interpret. They say, specifically, that “Focused Locations is a location type that represents the relevant locations of events or incidents described in a target message. Focused Locations are identified by selecting locations of interest from Locations in Text. ” (Ikawa et al., 2013, p. 1014). This description implies that focused locations are a subset of locations in text and can only be identified if named explicitly. The category does, however, seem to recognize the potential that inclusion of a proper place name in a tweet is not always an indication that the tweet is about that place (e.g., the tweet “I’m stuck in Chicago, but wish I could help recovery from the tornado in Moore, OK” is “about” Moore, OK and the reference to Chicago, while perhaps important if it is linked to other impacts of storms that canceled flights at the airport, is a secondary reference. Thus, the “focused locations” category seems to be a sub-category of locations in text and it does not seem to include the potential of the inferred locations that our “about” category does.

Given that several authors have already demonstrated some potential to link tweets to locations even when they do not mention the locations explicitly (see below), we contend that our 3-element typology is more complete and less ambiguous than the one proposed by Ikawa, et al (Ikawa et al., 2013). However, it is also clear that the *about* category can be further distinguished on the basis of whether: (a) the about location is mentioned explicitly (e.g., “the hurricane hit New Orleans head on and devastated the city” versus “I can’t believe the damage I’m seeing everywhere” – with the location of the latter implicit based on knowledge that everyone is talking about a particular story that hit a particular city), and (b) explicitly mentioned places are a primary or secondary focus (e.g., “the fighting here in Damascus is worse than I ever saw back when I was in Bagdad”; with “Damascus as the place that the tweet is really about and “Bagdad” is a secondary places used as a comparison).

There are a wide range of open questions about compiling place data associated with tweets and other social media posts and about leveraging those data to create useful information. In this section, we briefly review research focused on: (a) determining the characteristics of place data available in Twitter, (b) determining the places tweets are from, (c) determining the places tweets are about, (d) developing visual interface tools to support place-based Twitter analytics, and (e) applying analytical methods to location data associated with tweets to understand human activity and provide situational awareness.

2.1.1 Determining characteristics of place in Twitter

Several studies have reported on specific characteristics of place information that can be obtained from Twitter. Recent research has considered the frequency and validity for use of different place references in tweets (i.e., coordinates, place names supplied as location for individual tweets, and entries in the “location” field within user profiles) (Leetaru et al., 2013).

Many authors have published statistics suggesting that between 1 and 2% of tweets contain geolocation information that the user specifically opted to allow (a recent analysis of 1.5 billion tweets that found 1.6% with exact location, 1.4% with place information (typically named places), and 2% with geolocation overall, Leetaru et al., 2013). The percentages reported, however, are probably not consistent across tweet topics. In one example, Guy, et al (2010) find that 15% of the tweets containing one of the terms *earthquake*, *quake* and *tsunami* (in several languages) include geolocation (i.e., a *from* location in the form of coordinates or a place name). Below, we explore differences in propensity to include geolocation for tweets on other topics. In addition, the percentage varies geographically; Leetaru, et al (2013) report a range across the top 20 cities in the world of 2.86% in Jakarta followed by 2.65% in New York to as low as 0.91, 0.9, 0.88, and 0.85 for Dallas, Manila, Brussels, and Tokyo, respectively.

Hecht, et. al (2011) explored use of the “Location” field in Twitter profiles for over 5 million distinct users who issued approximately 32 million tweets (in June 2010). They found that approximately one third of users did not insert a real location in the field. Of the two thirds who did include a real location, few used locations more precise than a city. Hecht, et al. also analyzed tweet context and found that they could predict a user’s country and state from the contents of their tweets (at levels well above chance).

2.1.2 Determining the places tweets are from

Determining the place that tweets are from can be treated as a simple problem of reading the COORDINATES, GEO, and PLACE fields provided in metadata for each tweet and either using coordinates provided directly or converting named places in the place field into coordinates. Both the COORDINATES and GEO fields contain coordinates (if one has content, so does the other), but Twitter reports coordinates as "lat, lon" in GEO field, and as "lon, lat" in COORDINATES field (the former is a now "deprecated" field that Twitter continues to provide for compatibility reasons).

When coordinates are absent, there is also the potential to use various reasoning methods applied to the tweet text, metadata, and related information to estimate where the tweet may have been from. This has been approached by many others on a per tweet basis, attempting to leverage other references to location within the tweet metadata or text (e.g., Cheng et al., 2010; Gonzalez et al., 2012; Leetaru et al., 2013) and by using a range of context-based strategies to infer location of the user; these include combining content with tweeting behavior (Mahmud et al., 2012), content and social interactions (e.g., Chandra et al., 2011; Davis et al., 2011), and inferring current location from past behavior (e.g., Li et al., 2011). A missing perspective in much of this work is any overarching conceptualization of the place-time-concept context within which tweets are posted and thus can be interpreted. In related (non-social media) research, Tomaszewski and MacEachren (2010; 2012) present a conceptual framework and methods for applying geo-historical context to the task of foraging and sensemaking about place-based crisis situations using text-based news reports as the source; this framework has the potential to connect approaches to using context to infer location in tweets.

Leetaru, et al (2013), in recent work targeted at determining location of the tweeter when the tweet was posted (for the 98% of tweets lacking precise geolocation) focused on the profile location field as a possible source of relevant information. They found that approximately 1% of tweets contain coordinates in that field. They also found that (at 1x1 degree grid precision) the information in the profile location field (place name, which is much more common, or coordinates) matches with the precise coordinates in the coordinate field for 24% of tweets (at r=.52). They suggest from this that using profile location as an estimate of tweeter location (for the 98% of tweets that users have not turned location on for) has a 25% chance of being accurate This interpretation, however, assumes that users who put accurate and relatively precise information about a location in their profile are no more likely than other users to also turn exact location on; thus the assumption ignores an inherent bias in their test. We propose that those who are willing to provide precise location in their profile are probably less concerned with their own privacy, thus more likely to turn location on. To assess this, we compared statistics for our two databases detailed below (one with tweets collected based on mentions of place and the other collected based on having geolocated tweets). Tweets with geolocation were found to have non-null content in their profile location field 67% of the time and those in our larger database (with 1.5% of tweets geolocated) were found to have non-null content in their location profile just 57% of the time. Thus, if the objective is to estimate location from which tweets were posted without geolocation included, the success rate is likely to be less than the relatively low 25% that Leetaru, et al (2013) predict, since their analysis used only geolocated tweets.

2.1.3 Determining the places tweets are about

Tweets are obviously broadcast *from* a place, but in many situations, the place they are *about* is more important than the location of the person posting the tweet. There is a rapidly growing interest in adapting named entity extraction (NER, e.g., Liu et al., 2013), geographic information retrieval (GIR, e.g., Lieberman and Samet, 2012), and related geographic disambiguation strategies (e.g., Gelernter and Mushegian, 2011) to the challenge of determining the place(s) that tweets are about (or otherwise relevant to). This research can be divided into (a) work focused on recognizing explicit place references in the text of tweets (e.g., determining whether “Columbus” is a location, person, or component of an organization reference), (b) studies focused on disambiguating and geolocating the place referenced (e.g., once “Columbus” is found to be a location, determining whether it is Columbus, Ohio, Columbus, Georgia, or one of the other 166 populated places in the world called Columbus), and (c) efforts to develop integrated systems combining both processes.

Since this topic is only a small component of the research reported here, we do not attempt to review this research comprehensively. Representative recent research related to the NER component includes work by Lingad, et al (2013) directed to retraining standard NER methods to work with tweets; by Liu, et al (2013) to address the problem of normalization of named entities to their unambiguous canonical forms (this work is not place-focused); and by Sixto, et al (2013) to address the challenge of slang and abbreviations in tweets. For additional NER examples and related geographic disambiguation methods, we encourage interested readers to consult a recent paper by Gelernter and Balaji (2013) who provide an overview of recent research work as a background for introducing their end-to-end system. In their own research, they have developed and demonstrated strategies that produce substantially better results than standard NER tools. Their approach combines heuristics, machine learning, and open-source named-entity recognition software to achieve an average F-statistic of 0.90 for identifying place references (including streets, buildings, toponyms, and place abbreviations).

2.1.4 Developing visual interfaces to support place-based Twitter analytics

Multiple recent research efforts have focused on development of visual analytics methods and tools designed to enable geoinformation foraging and sensemaking with the very large, unstructured, and streaming data that is generated by Twitter and other social media sources. Twitter has been a particular focus of much research because the data is primarily public (in contrast to Facebook, LinkedIn, etc) and worldwide coverage is now substantial and continuing to grow. Publically accessible photo sharing sites have also attracted attention. We focus here on efforts to leverage Twitter, but related examples of research focused on photo sites are also relevant (Andrienko et al., 2010; Mirkovic et al., 2012; Zheng et al., 2011).

Dörk, et al (2010) were among the first to design and implement web-based, multiview visualization methods targeted toward monitoring the Twitter stream to understand people and topics of discussion. Their work focused on the combination of visual tools and system architecture to enable monitoring of the live, continually updating Twitter stream, which they demonstrated in real time in their 2010 conference presentation during which #visweek tweets appeared live in their interface as they were posted. While earlier work by Dörk and colleagues (2008) included geographic representations in their multi-view visualization strategy, their “Backchannel” system was non-geographic and focused on monitoring public reactions to events; they used the conferences they presented as test cases in which the backchannel discussion happened as the conference presentations and other events proceeded. Diakopoulos, et al (2010) presented a related system focused more specifically on support for journalists attempting to monitor public reaction to events (e.g., the State of the Union address); their system applied computational methods to filter data for uniqueness and to reveal user sentiment.

Visualization research that focuses on place as a component of information derived from Twitter data builds on a range of past work in Information Visualization and Geovisualization. In our own prior work, we implemented related web-based, multi-view visualization methods with a particular focus on place-based sensemaking and situational awareness for crisis events (see: MacEachren et al., 2011a; MacEachren et al., 2011b). That work underpins the research reported here and it directed particular attention to the distinction between *from* and *about* information contained in tweets. Focusing on cartographic display, Field and O'Brien (2010) introduce a concept they term "cartoblography" that they define as "a framework for mapping the spatial context of micro-blogging". As part of their approach, they propose spiral, spatio-temporal timeline visualizations of tweets that provide access to place-anchored commentary as it develops over time, putting emphasis on the most recent posts while providing a representation of and access to the past. Research grounded in visual analytics has integrated increasingly sophisticated computational methods accessed through flexible visual interfaces, producing systems focused on extracting meaningful and actionable information from tweets (e.g., Jie et al., 2012; Morstatter et al., 2013; Thom et al., 2012). In one recent study, Chae et al (2012) present a system that processes all geolocated tweets for the globe in real time with a focus on supporting analysis of abnormal events. The system uses topic modeling to extract topics from the geolocated Twitter stream and then applies abnormality estimation using *Seasonal Trend Decomposition*. They demonstrate the potential of the interactive system with case studies focused on a shooting event at a high school in Chardon, OH and the Occupy Wall Street protests.

2.1.5 Understanding human activity and supporting situational awareness

In addition to research on new analytical methods, there has been a wide range of research on understanding human activity and supporting situational awareness using Twitter and related social media. Some of the earliest work focused on determining what can be learned from Twitter as a messy, unstructured data source, but one that potentially has finer grained geographical information than most other open data (as long as relevant information can be extracted from the noise). Using manual analysis of tweets related to fire and flooding events, Vieweg et al (2010) were able to identify many on-topic tweets containing geolocated or place-specific content as well as on-topic situational updates. Tweeting behavior differed between the fire event and the flooding event (there were more geo-located tweets with the fire event than with the flooding event and, not surprisingly due to the timeframe, tweets related to flooding revealed more preparatory activity). In complementary research also targeted toward analysis of Twitter use in crisis situations (the Icelandic Volcano of 2010, in this case), Sreenivasan, et al (2011) find that tweets designed to "enlighten" are the most common category (defined as "users providing contextual information to better understand the situation"), with other important categories being status messages, problem understanding and factual data. In the context of a terrorist event, Oh, et al (2010) applied situational awareness theory to analyze tweets during the Mumbai terrorist attack in November, 2008; they found that "… 17.98 percent of posts contained situational information which can be helpful for the Mumbai terrorist group to make an operational decision of achieving their Anti-India political agenda." They also present complementary evidence that the terrorists actively monitored live media to enhance their own situational awareness.

The analyses highlighted above are largely manual, using simple data filtering to support human analysis. Other recent research has applied computational and visual analytics methods to explore place-based information extracted from Twitter more deeply. Crooks, et al (2012), analyze data from the Mineral, VA earthquake to demonstrate that Twitter can act as a distributed sensor system rivaling a planned physical sensor network in accuracy and with reduced cost and time to process information. Beyond use of Twitter as a sensing device, other work focuses on understanding human behavior and supplementing more traditional situational awareness methods. In one early study focused on tweeter behavior in crisis situations, Mendoza, et al (2010) analyzed tweets during the Chilean 2010 earthquake to determine the extent to which extracted information represented valid information versus baseless rumors. Their approach starts with an analysis of the social network of the community producing the tweets and they identified characteristics of how trending topics behave in the crisis situation and how they propagate through the network. Based on a small sample, they cite evidence that it may be possible to distinguish truth from rumor through analysis of propensity of individuals in the network to question the information (users question rumor more than valid information). More recently, Kent and Capello (2013) analyze use of Twitter during the 2012 Horsethief Canyon Fire in Wyoming and conclude that it is possible, using exploratory spatial regression analysis methods, to derive demographic characteristics for communities that relate to the likelihood that social media will generate meaningful data during a crisis event (with both nearness to the event and proportion of the population under 18 being positively related to the generation of meaningful data).

In a recent effort to go beyond exploration of data from a single source, Tsou, et al (2013) present a system that identifies and monitors spatial patterns for selected topics within publicly accessible web pages and public / semi-public social media. They demonstrate the approach and tools through application to analysis of the U.S. Presidential election. Representative results show maps of probability across the U.S. for hosting web pages supporting one or the other presidential candidate, analysis of relative proportion of tweets favoring each candidate by city, and spatial variation in tweet vocabulary.

2.2 Penn State insight into each place type

Here we summarize some results obtained about each place type in Twitter based on analysis of two sets of Twitter data: Geolocation Stream (using a spatial bounding box filter) and Keyword Stream (using a set of event-related keywords emphasizing crises). Before providing insights about each of the three place types (*from*, *affiliation*, and *about*), we describe and report statistics for the two data sets. This is then followed with a discussion of findings for each place type.

2.2.1 Geolocation Steam data set

The first data set (Geolocation Stream data set) is a small sample of 813,033 tweets collected using the Twitter streaming API with a spatial bounding box query that includes the conterminous U.S. and portions of Mexico and Canada (with a southwest corner of -124.7716944, 24.52083333 and a northeast corner of -66.94702778, 49.38447222). Data were collected for three days from May 7, 2013 to May 10, 2013. This data set includes all tweets within the bounding box that Twitter considers to be *from* somewhere within the bounding box during the time period, regardless of the content of those tweets. Tables 1a, 1b, and 1c below summarize key features of this data set.

One important aspect of the Place field in Twitter is that it contains an entry for almost all tweets having coordinates, not just for those that report geolocation without coordinates. When coordinates are included with a tweet, the Place field entry appears to represent a named place selected by the software platform used to post the geolocated tweet. Our Geolocation Stream data set (largely U.S. focused) includes 8,100,320 tweets with coordinates and all except 25,187 of these also had a Place ID. But, there are only 6,638,874 non-duplicate coordinates and 6,626,628 tweets from different coordinate locations that also contain Place IDs (12,687 of the tweets with non-duplicate coordinates lack Place IDs). Within this sample, we found 95,578 unique Place IDs to which the more than 6 million coordinates are mapped. Additional explanation of data in the tables is provided in relevant sub-sections below.

Table 1a: Statistics for the Geolocation Stream data set

tweet	# of all tweets (with coordinates)	8,106,244	*	
	# of tweets excluding duplicates	8,100,320	|tweet id|	This leaves out tweets with identical content based on tweet ID.
	# of tweets excluding duplicates	8,091,240	|user id & time & place id & coordinates|	This leaves out tweets judged identical by common user, time, place, and coordinates.
tweet	# of unique place id	93,578	|place id|	* There are tweets (25,187 out of unique tweets with coordinates 8,100,320) with coordinates, but without place id. e.g. 332941030641000448 -- it accounts for the difference between # of place id from tweet table and place table.
	# of unique coordinates	6,638,874	|coordinates|	
	# of unique coordinates & places	6,626,628	|place id & coordinates|	* 12,687 coordinates don't have |place id|. So, **6,626,187** unique coordinates & place types have |place id|.
place	# of unique place id	93,577	|place id|	
	# of unique place id or country	93,597	|place id or p_country|	
	# of unique place id or type	93,597	|place id or p_placetype|	* Each |place id| has only one type of places. * Some |place id|s don't have its |place full name|. e.g. place id = '50ff257b9fe4a92f' ** Some |place id|s have multiple full names. But, there is no semantic difference among multiple full names -- those indicate just one place. This implies that we can just use |place id| field to detect unique places.
	# of unique place id or full name	93,764	|place id or p_fullname|	
	# of unique place id or country or type or full name	93,781	|place id or p_country or p_placetype or p_fullname|	

	# of unique place id & country or type or full name or address	93,781	|place id or p_country or p_placetype or p_fullname or p_street|
user	# of unique users	573,909	|user id|
	# of unique username	574,774	|username|
	# of unique user & username	574,774	|user id & username|

Table 1b. Statistics for Place IDs represented in the sample of unique coordinates

# of unique coordinates	6,638,874	100.00
# of unique coordinates **without** |place id|	12,687	0.19
# of unique coordinates **with** |place id|	6,626,187	99.81
# of unique coordinates with |place id| & **|city type|**	5,717,450	86.12
# of unique coordinates with |place id| & **|admin type|**	816,418	12.30
# of unique coordinates with |place id| & |**poi type|**	66,377	1.00
# of unique coordinates with |place id| & **|neighborhood type|**	19,693	0.30
# of unique coordinates with |place id| & **|country type|**	6,249	0.09

Table 1c. Top 20 Place IDs in the sample based on number of tweets with this ID; note that all are of the city or admin (state/province) types

1	Los Angeles, CA	120,973
2	Texas, US	93,980
3	Georgia, US	89,901
4	Ohio, US	74,392
5	Chicago, IL	73,261
6	Florida, US	71,795
7	South Carolina, US	70,701
8	Manhattan, NY	68,515
9	Philadelphia, PA	61,849
10	Houston, TX	61,031
11	New York, NY	57,627

12	San Antonio, TX	56,929
13	Dallas, TX	46,496
14	Toronto, Ontario	40,522
15	Maryland, US	38,243
16	California, US	37,867
17	Austin, TX	36,557
18	Pennsylvania, US	34,780
19	Boston, MA	34,215
20	Indiana, US	33,549

2.2.2 Keyword Stream data set

The second data collection (Keyword Stream data set) consists of 209,458,317 tweets collected for Jan. 1, 2013 – June 10, 2013 using the Twitter streaming API and keywords that focus on crisis events, public health, protests, and (more recently) airports. This data set includes all tweets returned by Twitter that match the keywords, whether or not they include geolocation in the form of coordinates or Place IDs (*from*) or place names within the text (*about*). As indicated in Table 2, about 10% of all tweets collected include place references in the text and about 1.5% include geolocation in the form of coordinate. Based on our separate analysis of geolocated tweets, we estimate that 9% of the geolocated tweets have only a Place ID with no coordinates (thus only about 0.1% of our entire database). We currently have 337 keywords in the set categorized as: 115 crisis/event related terms, 54 place names of particular interest (e.g., due to events that happened there), 115 airport-related terms included to focus on travel delays due to storms and other events, and 53 other terms related to non-crisis events in the news. Of the 337 keyword, 55 are hashtags; these include general terms used in many events (#volunteers, #curfew, #info) and others specific to particular places or events (#egypt, #newsburkinafaso, #jan25). Table 2a below provides summary statistics for this data set and Table 2b provides statistics for tweets containing each keyword plus one or both kinds of location (the denominator for each row in Table 2b is the total for tweets containing the respective keyword).

Table 2a: Statistics for Keyword Stream data set

Features	Frequency	%
ABOUT	21,827,739	10.42
FROM	3,140,585	1.50
"protest"	187,242	0.09
"tornado"	631,937	0.30
"fire"	466,119	0.22
"earthquake"	349,906	0.17
"flood"	535,910	0.26
"Miami"	196,813	0.09
"Boston"	359,737	0.17

Table 2b. Statistics for place references by keyword

"protest" & FROM	661	0.35
"tornado" & FROM	11,246	1.78
"fire" & FROM	5,820	1.25
"earthquake" & FROM	26,710	7.63
"flood" & FROM	22,554	4.21
"Miami" & FROM	2,809	1.43
"Boston" & FROM	3,704	1.03
"protest" with ABOUT	109,608	58.54
"tornado" with ABOUT	182,687	28.91
"fire" with ABOUT	137,466	29.49
"earthquake" with ABOUT	156,299	44.67
"flood" with ABOUT	122,165	22.80
"protest" & FROM + ABOUT	301	0.16
"tornado" & FROM + ABOUT	2,451	0.39
"fire" & FROM + ABOUT	1,675	0.36
"earthquake" & FROM + ABOUT	21,268	6.08
"flood" & FROM + ABOUT	7,751	1.45

2.2.3 Place the tweet is from

Using our Keyword Stream data set, we have calculated selected statistics related to how often and for what kinds of topics geolocation is assigned to tweets by users. As has been reported in multiple sources, the feature to assign *from* locations in Twitter is not widely used. Approximately 1.5% of the tweets in our database of 209 million tweets include a *from* location, thus they contain non-null entries in the COORDINATES, GEO, and/or PLACE metadata field (Table 1). This is in the 1-2% range typically reported. However, a preliminary analysis of variation in propensity to include geolocation based on event type or place that a tweet is about suggests that there may be quite large differences. For event type, we find a range from 0.35% with geolocation for tweets containing the term “protest” to 7.63% for tweets containing the term earthquake; the latter corresponds with previous findings that there is a higher than average proportion of tweets containing geolocation when earthquake is a topic (Guy et al., 2010).

Using our Geolocation Stream data set, we analyzed what can and cannot be determined about the place tweets are from when users elect to turn location on. A primary focus of this analysis is on understanding the quality and variability of location data provided. As noted, this data set consists entirely of tweets that Twitter has determined to be from locations within a bounding box that primarily covers the conterminous United States, with portions of Canada and Mexico. Thus, results may not represent locations outside the U.S. well.

As noted above, when tweets include geolocation in the form of coordinates, they almost always include an entry in the Place field. This field is the only location information for the 9% of tweets with geolocation that lack coordinates. Table 3, summarizes multiple characteristics of the Place field in tweets. It includes, or provides access to, a rich set of information beyond simple coordinates.

Table 3. Characteristics of metadata associated with the "Places" field of tweets

Category	Field	Type	Description
Places	id	String	Twitter's ID representing this place. Note that this is represented as a string, not an integer. Example: "id":"7238f93a3e899af6"
	name	String	Short human-readable representation of the place's name. Example: "name":"Paris"
	full_name	String	Full human-readable representation of the place's name. Example: "full_name":"Paris, Paris"
	place_type	String	The type of location represented by this place. There are basically five types of places: **city, admin, country, neighborhood, poi**. Example: "place_type":"city"
	country	String	Name of the country containing this place. Example: "country":"France"
	country_code	String	Shortened country code representing the country containing this place. Example: "country_code":"FR"
	url	String	URL representing the location of additional place metadata for this place. Example: "url":"http://api.twitter.com/1/geo/id/7238f93a3e899af6.json"
	bounding_box	Object	A bounding box of coordinates which encloses this place.
		Field	Description
		coordinates	**Array of Array of Array of Float**. A series of longitude and latitude points, defining a box which will contain the Place entity this bounding box is related to. Each point is an array in the form of [longitude, latitude]. Points are grouped into an array per bounding box. Bounding box arrays are wrapped in one additional array to be compatible with the polygon notation. Example: "coordinates":[[[2.2241006,48.8155414], [2.4699099,48.8155414], [2.4699099,48.9021461], [2.2241006,48.9021461]]]
		type	**String.** The type of data encoded in the coordinates property. This will be "Polygon" for bounding boxes. Example: "type":"Polygon"
	attributes	Object	Contains a hash of variant information about the place. Place Attributes are metadata about places and allow any user or application to add arbitrary metadata to a place. An attribute is a key-value pair of

			arbitrary strings, but with some conventions. Keys can be no longer than 140 characters in length. Values are **unicode strings** and are restricted to 2000 characters.
			Example:
			"attributes": {
			"street_address": "795 Folsom St",
			"623:id": "210176",
			"twitter": "twitter"
			}

		Well-known attributes: There are a number of well-known place attributes which may, or may not exist in the returned data. These attributes are provided when the place was created in the Twitter places database.	
		Keys	Description
		street_address	Example: "street_address": "795 Folsom St"
		postal_code	in the preferred local format for the place
		phone	in the preferred local format for the place, include long distance code
		twitter	twitter screen-name, without @
		url	official/canonical URL for place
		app:id	An ID or comma separated list of IDs representing the place in the applications place database. Example: 174368:id: "202500033005894", 174368:admin_order_id: "FRA:11::::::75:75056", 189390:id: "washington-dc", 162772:place_id: "1150000", 162772:pop100: "572059"

	contained_within	Places	A place which encloses this place. This Places object can be the city the place is in, the administrative region the place is in, or so forth.
			Example:
			"contained_within": [
			{
			"name": "San Francisco",
			"country": "United States",
			"country_code": "US",
			"attributes": {
			},
			"url": "http://api.twitter.com/1/geo/id/5a110d312052166f.json",
			"bounding_box": {
			"coordinates": [
			[[-122.51368188, 37.70813196],
			[-122.35845384, 37.70813196],
			[-122.35845384, 37.83245301],
			[-122.51368188, 37.83245301]]
			],
			"type": "Polygon"
			},
			"id": "5a110d312052166f",
			"full_name": "San Francisco, CA",
			"place_type": "city"
			}
			]

The Place IDs (stored as an alphanumeric value) also have text labels; in some cases more than one similar label is assigned to the same Place ID (e.g., “Subway, Houston” versus “Subway (Galleria Area), Houston”). The Place IDs include multiple feature types (country, administrative region including states, city, neighborhood, and POI). City is by far the most common type and country type the least common, see Table 1b. Many tweets may be linked to any particular coordinate location. In our sample of more than 6 million tweets with unique coordinates, 100,005 have 4 or more tweets posted from the coordinate location. The maximum tweets from a single coordinate location in the data set is for a location that maps to the Place ID for New York, NY; 1594 tweets share this coordinate location. The New York, NY Place ID, however, is associated with 57,627 tweets. It is clear that many different coordinate locations map into the one Place ID for New York, NY. The inverse also happens, specific coordinate pairs can be linked to more than one Place ID (not in the same tweet, but from different tweets that have the some coordinate location. As a step toward understanding the geographic characteristics of relationships between coordinates and Place IDs, we mapped those for all coordinates with 4 or more tweets linked to this Place ID (979 tweets). The map (Figure 2) illustrates a clustered distribution in which most places with a New York, NY Place ID are in Brooklyn or the Bronx and few are in Queens, Staten Island, or Manhattan. The latter three have their own Place IDs while the former two do not. In addition, the map also illustrates that there are a smaller number of locations outside the city bounds that are assigned to the city. Our interpretation (not confirmed at this point) is that this spatial variation results from the many different applications (e.g., for mobile devices) that have the ability to be used for posting tweets.

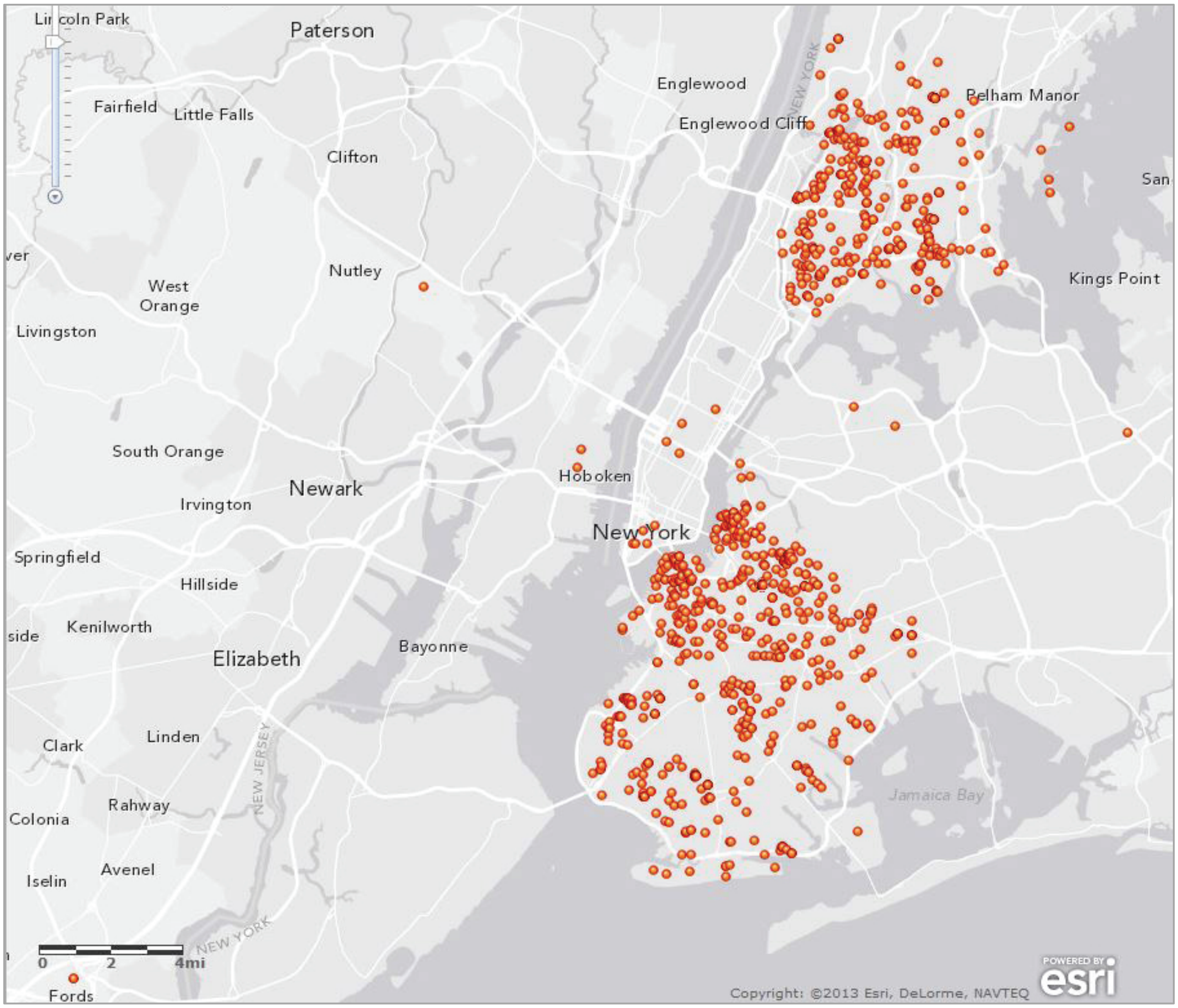


Figure 2: Locations from which multiple tweets with a Place ID for New York, NY were posted (locations shown had 4 or more tweets with identical coordinates).

Beyond *city type* Place IDs (the most frequent), those for *neighborhoods* and *POIs* (while fewer in number) have the potential to be used as a component of context that improves local (within city) place entity recognition and geographic disambiguation applied to the text of the tweets (since the locations that a user mentions have a somewhat higher probability to be near the location they are at than distant locations). Another interesting aspect of the Place field in Twitter metadata for geolocated tweets is that the 'url' of places provides additional geo-place attributes, particularly about the field of 'attribute' and 'contained_within'.

2.2.4 *Place affiliation:*

In addition to analyzing the two data sets collected, we have also analyzed a small sample of data (500 tweets sampled from our Keyword Stream data set) contained within the location field of user profiles. Analysis of this sample found the following:

- for 59% of the entries in the sample, the user profile includes some type of location reference; subsequent counts on the 20 million tweets containing recognized places in our Keyword Stream data set and on the full Geolocation Stream data set found 57% and 67% with non-null values in the location field. The latter suggests that individuals willing to provide the location of their tweets are also more likely to include an entry in the location field of their profile.
- the large majority of those in the 500 tweet sample with a non-null entry in the location field have a geographic location (32 in a sample of 500 are clearly not places -- thus approximately 53% have a location listed that is a real location)
- for the small proportion of non-locations, most should be relatively easy to separate from real locations, e.g., Back from the Void, Behind you, Quantum Leaping, The space between spaces
- when a geographic reference is provided, we identified several categories of challenges to extraction of meaningful information; these include fields with:
    - multiple places (e.g., Charlotte,Chicago,Dayton)
    - implied transit between places (e.g., from California to Arizona; ATL-NYC)
    - nicknames (e.g., Beantown)
    - qualifiers, prefixes, suffixes (e.g., Faretotheham ..unfortunately; iPhone: 38.894707,-77.027260; Sunny FL)
    - syntax, spelling problems, and apparent efforts to prevent automatic extraction (Caracas\Venezuela; Edmonton, Alberta!; Canadaaa; n e w y o r k)
    - non-places + places (e.g., Ki's Empire (Bogota D.C.) -- "Bogota D.C." is a real city);
    - vague locations, some of which may be tricky to recognize as well as to locate (e.g., Southern Illinois; Midlands, England; Near Manchester, NH)
    - combinations of the above (e.g., Willamette Valley Worldwide)

2.2.5 *The place tweets are about*

In contrast to the 1-2% average for tweets with geolocation, we find that approximately 10% of the tweets in our Keyword Stream data set contain *about* features, thus they have identifiable named entities that are recognized as locations. Preliminary analysis using about 200 hand-coded tweets suggests that this percentage is a substantial underestimate. Analysis of the GATE entity extractor we have used with this test set finds that only 64% of tweets with named places have all of the named places recognized correctly as locations. This suggests that the 10% found may represent only about 2/3 of the tweets that have places mentioned in their text; thus 15% is probably a better estimate for tweets that contain named places, if the entity extractor can be trained to identify them more accurately. The propensity *about* information reported here should only be interpreted as reflecting tweets relevant to crisis and related events since the database of tweets analyzed is compiled utilizing keywords designed to represent these topics.

Thus, an unanswered question is whether there are significant differences in frequency of *about* references based on location in the world, tweet topic, or other factors. As a preliminary step toward answering this question, we compare (Table 2) the proportion of tweets containing specific event terms that also have identified place entities. We find that tweets with "protest" top this list with 58.5% (after being at the bottom of the list of tweets with geolocation). For the other natural disaster related topics, tweets with "earthquake" are most likely to reference named places (44.7%) while those with "flood" are at the bottom of the list (22.8%).

3 Place in Twitter Query and Visualization

This research focuses on deriving, accessing, and visualizing place-based information from open media (Twitter in particular). In this component of the research, we addressed the following objectives: (a) to extend the characteristics of place for which SensePlace2 enables access and enhance the place-specific query capabilities of SensePlace2 that supports the access and (b) to implement visual methods that support exploration of the additional place characteristics and query results.

3.1 Enabling access to place in Twitter

In relation to characteristics of place, the research reported here focused on adding capabilities to use additional place information beyond the existing ability of SensePlace2 to depict the geolocation that tweets are from (when this is indicated) and the place they are about (relying upon named entity recognition, geographic disambiguation, and geocoding tools developed in complementary research funded by the Department of Homeland Security). Specifically, the capability was added to support locations included in user profiles and locations highlighted through hashtags (e.g., using the #loc hashtag advocated by the Tweak-the-Tweet project and/or explicit place name hashtags such as #Boston). Both are now included in the Lucene-Solr index that underpins SensePlace2 (see section 4 for details).

For querying, several enhancements were implemented. SensePlace2 now supports a bounding box spatial query that selects tweets *about* any place inside the box or tweets *from* any place inside the box depending on whether the user has restricted the query to *from* locations or not. This complements the existing point-based spatial query that accesses the 1000 most relevant tweets weighted by distance from the user-specified point. In addition, hashtags are now interpreted as exact matches; this makes it possible to query specifically for hashtags that are a place name (e.g., #Boston) and #loc hashtags advocated by the Tweak-the-Tweet project. When additional query terms are related to crisis events (e.g., flood, hurricane, bombing, etc.), most instances of #loc are used to specify a location, which follows the hashtag. In ordinary tweets, #loc is used frequently to indicate things other than location.

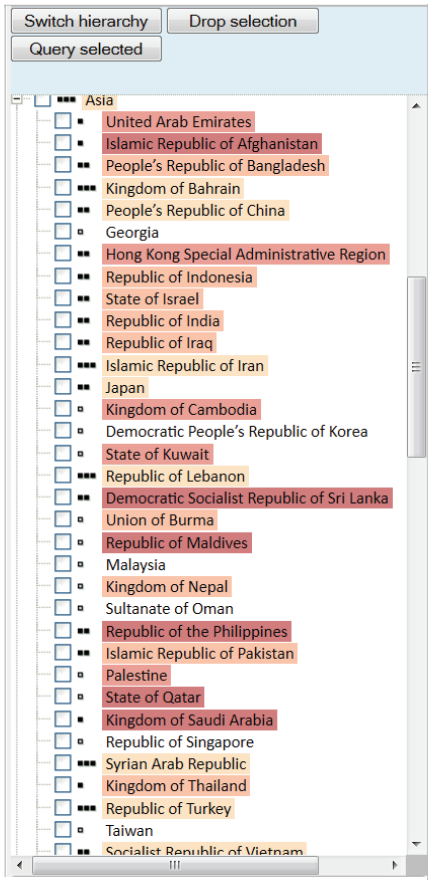
3.2 Visual support for place-based analysis

Work to provide visual support for place-based analysis (in support of Task 1) included: extension to existing methods within SensePlace2 (a place-tree hierarchy, place-focused tag clouds, and place co-referencing on the map); implementation of a new visual method within SensePlace2 (a dynamic co-occurrence matrix); and initial work on strategies for analysis of movement from geolocated tweets. Each is detailed below.

3.2.1 Place Tree

The Place Tree Hierarchy was introduced to SensePlace2 as part of previous work during Stage 1 of this project (Figure 3). The Place Tree is built to follow the structure of the GeoNames place hierarchy and is currently populated down to the country level. Each of the nodes in the hierarchy is colored according to the number of matches the given query has in the entire database, whereas the stacked black dots represent the number of matches in the top 1000 tweets.

The most important change in the Place Tree Hierarchy is that it now works within the overall component-coordination model detailed below. From a user perspective, a key benefit is more consistent visual coordination between this component and others (e.g., when a user clicks the check box next to a country in the Hierarchy, tweets matching the current query parameters associated with that country are moved to the top of the Tweet list). In addition, the method to support how the place-tree hierarchy is drawn in the UI was re-written to take advantage of system architecture changes detailed below. The component now uses a "timed code" technique – it is now drawn by element in small batches, and a special bit of code makes sure these batches do not freeze the UI for more than 50ms. Thus, the UI stays fully interactive during hierarchy tree initialization. This is most obvious when the user accesses the "switch hierarchy" button.



□ ■■■ Asia
□ ■ United Arab Emirates
□ ■ Islamic Republic of Afghanistan
□ ■■ People's Republic of Bangladesh
□ ■■■ Kingdom of Bahrain
□ ■■ People's Republic of China
□ ▫ Georgia
□ ■■ Hong Kong Special Administrative Region
□ ■■ Republic of Indonesia
□ ■■ State of Israel
□ ■■ Republic of India
□ ■■ Republic of Iraq
□ ■■■ Islamic Republic of Iran
□ ■■ Japan
□ ▫ Kingdom of Cambodia
□ ▫ Democratic People's Republic of Korea
□ ▫ State of Kuwait
□ ■■■ Republic of Lebanon
□ ■■ Democratic Socialist Republic of Sri Lanka
□ ▫ Union of Burma
□ ▫ Republic of Maldives
□ ▫ Malaysia
□ ▫ Kingdom of Nepal
□ ▫ Sultanate of Oman
□ ■■ Republic of the Philippines
□ ■■ Islamic Republic of Pakistan
□ ▫ Palestine
□ ▫ State of Qatar
□ ■ Kingdom of Saudi Arabia
□ ▫ Republic of Singapore
□ ■■■ Syrian Arab Republic
□ ■ Kingdom of Thailand
□ ■■■ Republic of Turkey
□ ▫ Taiwan
□ ■■ Socialist Republic of Vietnam

Figure 3. Place-Tree Hierarchy for a query on "Protest" for June, 2013, scrolled to show results for Asia. Bahrain, Iran, Lebanon, Syria, and Turkey are all in the top third of places for the 1000 most relevant tweets, while Afghanistan, Sri Lanka, are in the top tertile in the full dataset, but not among the most relevant tweets.

3.2.2 Place Clouds

Leveraging the new component-based architecture for SensePlace2 (discussed in section 4.1 below), a second Place Cloud has been added to the interface. One of the place clouds depicts results for all locations in the database that match the query (the Overview Locations tab) and the other depicts all locations in the 1000 most relevant tweets that match the query (the Relevant-Tweet Locations tab). Thus, the user can quickly determine whether the boosting methods applied have emphasized tweets relevant to particular places in comparison to the full set of tweets matching the query entered (Figure 4). The mechanism used to populate the place clouds has been modified from the version reported in Stage 1 as well. Each Place Cloud now aggregates locations based on the toponyms associated with them rather than by their GeoNames IDs. This addresses the fact that several GeoNames IDs may be associated with what is conceptually (for most users) the same place. Keeping them independent resulted in low counts for some important places.

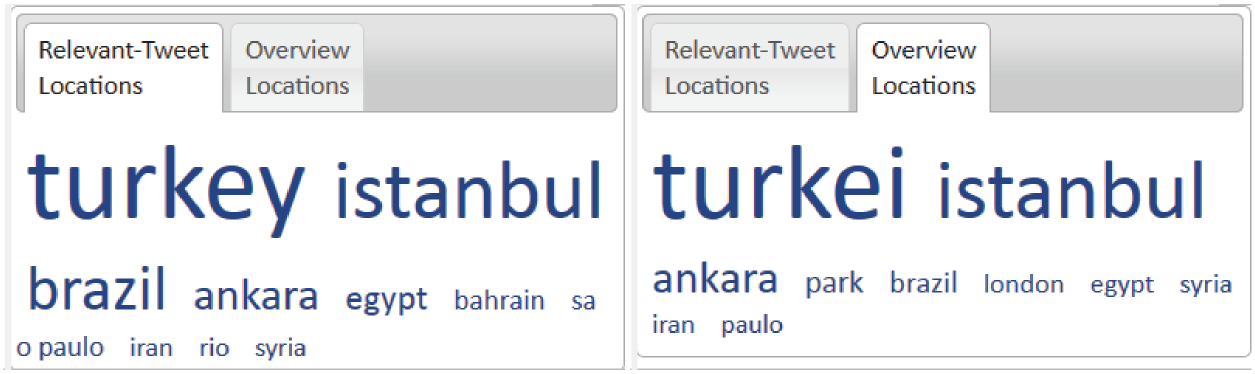
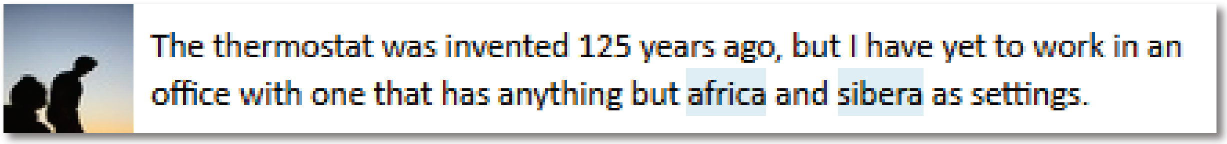


Figure 4. Place Clouds displaying the more frequently mentioned places in the 1000 most relevant tweets and the most frequently mentioned places in the overview; both represent a query on "Protest" for the month of June, 2013.

3.2.3 Place Co-Referencing on Map

A third strategy to understand place supported by SensePlace2 focuses on interconnections among places. The current implementation is directed to instances of joint reference to more than one location within individual tweets; only the 1000 most relevant tweets matching the current query (thus those accessible in the client application) are considered in calculating co-reference.

The basis for doing co-reference analysis is that tweets frequently refer to multiple locations at the same time (Figure 5). Thus, for each location, a list of co-occurring locations can be built. The mechanism used to build this list can be described as follows. Each tweet in the set of 1000 is treated as a collection of locations. Using this information, a list of unique locations is built. For each unique location in this list, we scan through the list of 1000 tweets again and make a record if and when the unique location under consideration is mentioned along with any other location.

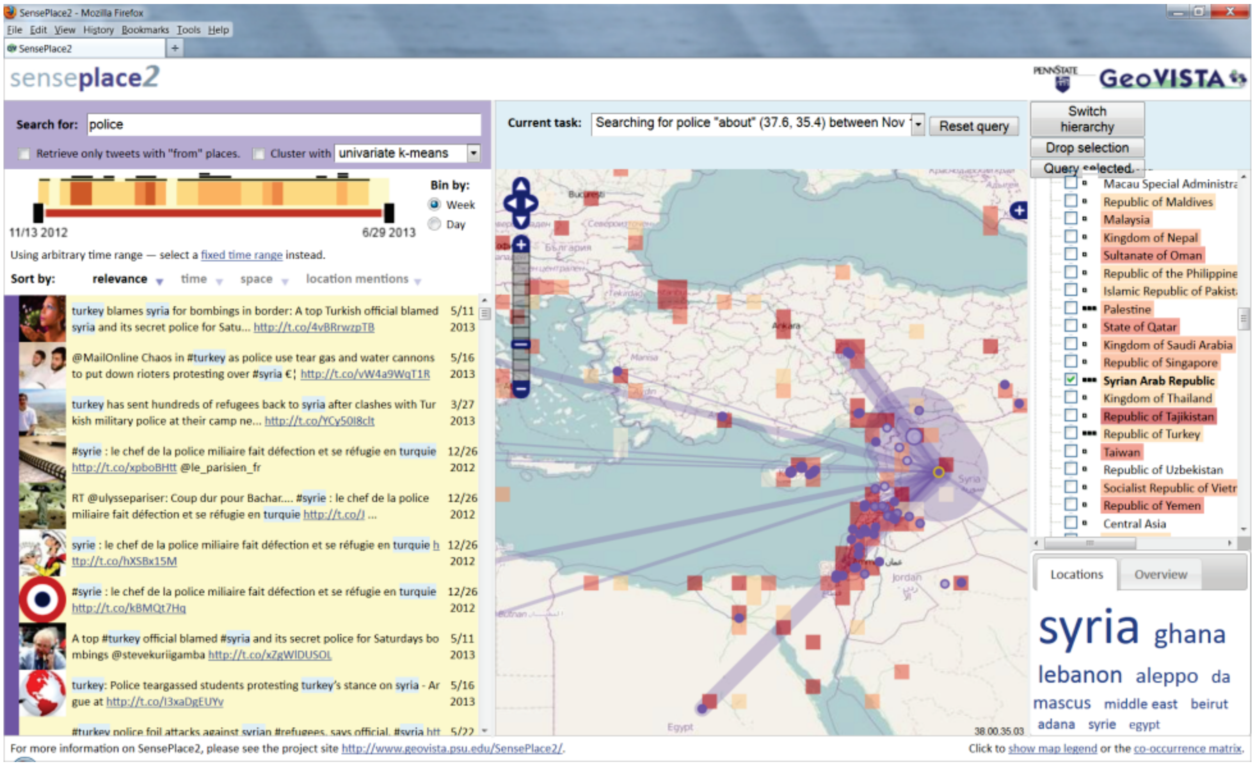


The thermostat was invented 125 years ago, but I have yet to work in an office with one that has anything but africa and sibera as settings.

Figure 5: An example tweet with two place names recognized and highlighted

Both *about* and *from* locations are included in this process, but since *about* locations are more frequent, they are more prominent in the results. When present, a *from* location is not treated any differently than the *about* location, except for the map symbols used (purple for *about* and green for *from*). Given two types of locations (*from* and *about*), two types of co-reference are possible: (a) two or more *about* locations (as shown in the tweet below) (b) a *from* location plus at least one *about* location.

When a particular location is highlighted on a map, the list of co-occurring locations is retrieved and shown in the form of connecting lines between the original and the co-occurring locations, as shown in the figure 6 below. The width of the line depicts quintiles of frequency for connections (bold lines represent more connections).



Search for: police

□ Retrieve only tweets with "from" places. □ Cluster with univariate k-means

11/13 2012 6/29 2013

Using arbitrary time range — select a fixed time range instead.

Current task: Searching for police "about" (37.6, 35.4) between Nov

turkey blames syria for bombings in border: A top Turkish official blamed syria and its secret police for Satu... http://t.co/4vBRrwzpTB	5/11 2013
@MailOnline Chaos in #turkey as police use tear gas and water cannons to put down rioters protesting over #syria €¦ http://t.co/vW4a9WqT1R	5/16 2013
turkey has sent hundreds of refugees back to syria after clashes with Turkish military police at their camp ne... http://t.co/YCy50I8clt	3/27 2013
#syrie : le chef de la police miliaire fait défection et se réfugie en turquie http://t.co/xpboBHtt @le_parisien_fr	12/26 2012
RT @ulysseparizer: Coup dur pour Bachar.... #syrie : le chef de la police miliaire fait défection et se réfugie en turquie http://t.co/J ...	12/26 2012
syrie : le chef de la police miliaire fait défection et se réfugie en turquie http://t.co/hXSBx15M	12/26 2012
#syrie : le chef de la police miliaire fait défection et se réfugie en turquie http://t.co/kBMQt7Hq	12/26 2012
A top #turkey official blamed #syria and its secret police for Saturdays bombings @stevekuriigamba http://t.co/xZgWIDUSOL	5/11 2013
turkey: Police teargassed students protesting turkey's stance on syria - Argue at http://t.co/I3xaDgEUYv	5/16 2013
#turkey police foil attacks against syrian #refugees, says official. #syria htt	5/22

□ ▪ Macau Special Administra
□ ▪ Republic of Maldives
□ ▪ Malaysia
□ ▪ Kingdom of Nepal
□ ▪ Sultanate of Oman
□ ▪ Republic of the Philippine
□ ▪ Islamic Republic of Pakist
□ ▪▪▪ Palestine
□ ▪ State of Qatar
□ ▪ Kingdom of Saudi Arabia
□ ▪ Republic of Singapore
☑ ▪▪▪ **Syrian Arab Republic**
□ ▪ Kingdom of Thailand
□ ▪ Republic of Tajikistan
□ ▪▪▪ Republic of Turkey
□ ▪ Taiwan
□ ▪ Republic of Uzbekistan
□ ▪ Socialist Republic of Vietr
□ ▪ Republic of Yemen
□ ▪ Central Asia

For more information on SensePlace2, please see the project site http://www.geovista.psu.edu/SensePlace2/.

Click to show map legend or the co-occurrence matrix.

Figure 6: An example of the co-reference function on the SensePlace2 map; results shown are based on a spatial point location query for "police" with the point in Syria

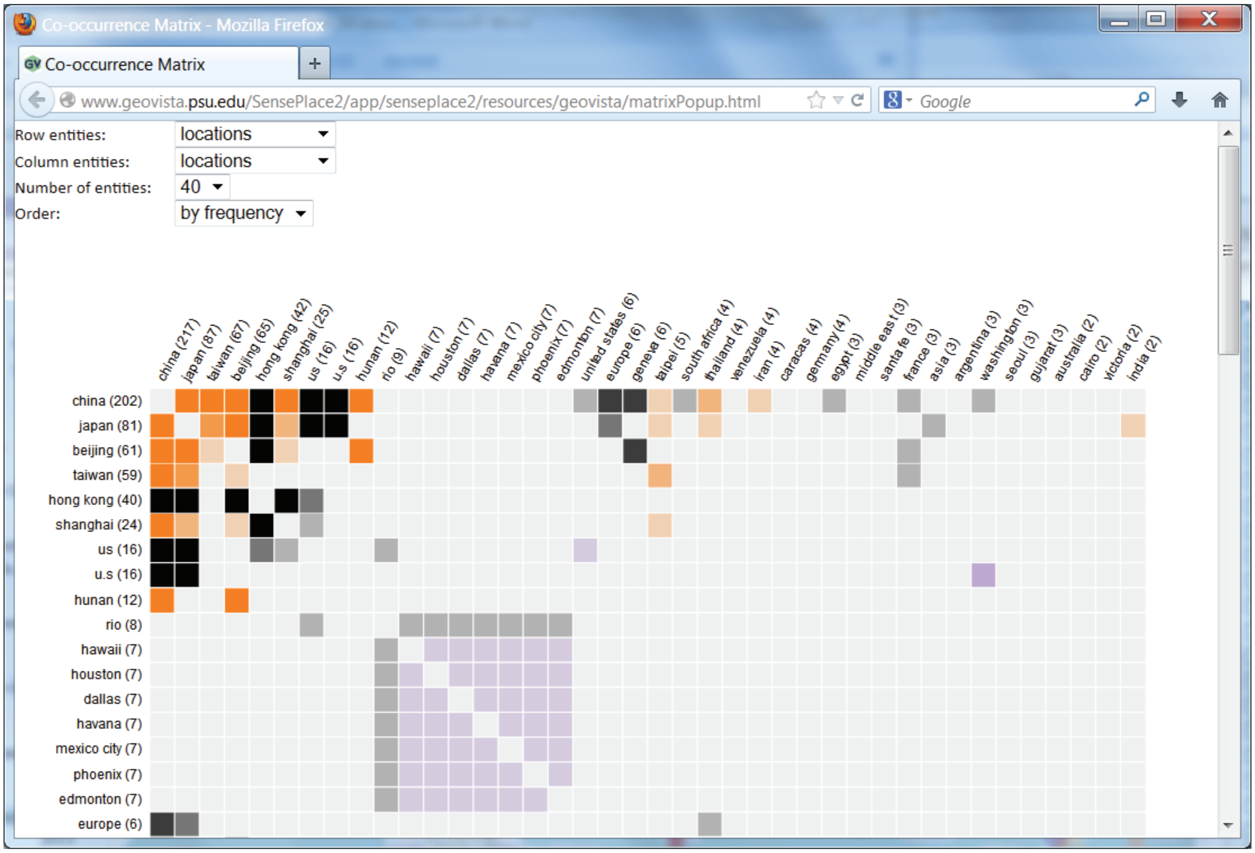
Connecting lines are drawn both for *from* and *about* locations, and are colored purple when both locations at the ends of the line are of *about* kind, and green when either of the two is of *from* type. Thus, for green connecting lines to be drawn, something has to co-occur with a *from* location, which would be at least one *about* location. This connection is treated as bi-directional. As a result, if you hover over a *from* location, you will see links to all of the *about* locations mentioned in the tweets that came from this location. If you hover over any of the *about* locations, you will see links to all other *from* and *about* locations that co-occur with the location selected in the 1000 tweets.

3.2.4 Place Co-Reference Matrix

In the single tweet image above, references to "africa" and "siberia" co-occur. The co-occurrence matrix provides a feature-based method to visualize this kind of relationship as well as many other kinds of joint occurrence. The default view (shown below) provides a depiction of location-location co-occurrence frequencies for a query on "flu". The user controls the number of columns to specify the N locations with the most mentions and the rows list every location mentioned in at least one of the 1000 most relevant tweets. Frequencies are grouped into tertiles, with dark fill representing the highest frequencies and light fill the lowest. Color hue currently depicts spatial homophily, specifically if the locations that co-occur are in the same continent, they are assigned a hue to represent that continent and if they are in different continents, the fill is gray.

The Co-occurrence Matrix has been designed to be flexible enough to match any pair or entities that can be determined for tweets (Figure 7). The current implementation allows users to match location (as detailed above, this includes both *from* and *about* locations if they exist) with: location (*from* and *about*), country (thus reference to any place within the country), continent (reference to any place within the continent), users, user profile location (when available), hashtags, and day of the week. Although our focus here is on place-based analysis, it is important to note that the Co-occurrence Matrix does not restrict the user to location; any pair of features can be related (e.g., hashtags matched to day of the week, users matched to place they refer to, etc.).

The matrix can be sorted in three ways, alphabetically by name, by continent, and by frequency. The example below was sorted by continent, then by frequency, grouping Asia and North America (the continents with the highest frequencies of “flu” mentions) at the top. Currently, all sorting options are available regardless of what types of entities are plotted on the co-reference matrix. However, sorting by continent (specifically, by GeoNames continent ID) does not work unless a row/column label can be assigned to a continent explicitly. The user is able to constrain the number of columns to be displayed (e.g., if they pick 40, then the 40 entities with the highest frequencies will be displayed). Scrolling allows all entities for the feature selected as row entities to be depicted.


Co-occurrence Matrix - Mozilla Firefox
Co-occurrence Matrix
www.geovista.psu.edu/SensePlace2/app/senseplace2/resources/geovista/matrixPopup.html
Google

Row entities: locations
Column entities: locations
Number of entities: 40
Order: by frequency

Figure 7: Location-Location Co-Occurrence Matrix results for mentions of “flu”

3.2.5 Movement Analysis

A potentially rich direction of future work is the exploration of movement data that can be built from tweets with explicit coordinate information. As an extension from Task 1 research reported here and methods used to construct the Geolocation Stream data set, a preliminary analysis of a sample dataset containing 9 million tweets identified a total of 7.5 million movement records. Although most of these records (about 80%) describe movement at short distances and at speeds below 1 MPH, there is ample data available to look into high-speed, long-distance movement as well.

Although movement analysis is not within the scope of our current USACE research, a promising future research objective related to applications of movement data from Twitter is classification of tweets by mode of transportation. Knowledge that particular tweets are associated with travel by car, by train or by plane can be treated as a semantically-rich variable that can be combined with a set of spatial and keyword-based queries. Outside of our current USACE effort, we are working on a prototype visualization environment that allows us to explore distance, time and speed dimensions of the dataset mentioned above.

4 System Architecture Enabling Dynamic Control, Query & Visualization

This section reports on development and implementation of extensions to our SensePlace2 application that underpin the advances in visual analytical methods and tools reported above. First, we outline the approach developed for browser-based interface component coordination. Second we describe the coordinator component implemented. Then, we outline some of the advantages of the approach. Finally, we summarize the introduction of Solr indexing that enables faceted query and increased performance for large data sets.

4.1 SensePlace2 User Interface Coordination Mechanism

Internally, the SensePlace2 user interface (UI) is composed using a number of components. Some of the more obvious components include the list of 1000 most relevant tweets, map, timeline with associated controls, place-clouds, and place-tree hierarchy. There are also several components that are less obvious or hidden in the UI (e.g., history widget, system status message, rank controls for tweets in the tweetlist, etc.). Our component-based approach makes software development and maintenance considerably easier and improves the overall stability of the application. Once individual components are put in place, it is possible to begin linking them together into a cohesive interactive visualization environment.

SensePlace2 UI components are interactive in two distinct ways. First, they respond to user actions (e.g., mouse hover or click) performed in the interface. When a user clicks a point symbol on a map, the point symbol is highlighted in a different color, the name of the location corresponding to that symbol is displayed in a miniature pop-up, and lines are drawn pointing to locations that are related to the one the user clicked on.

Second, they expose a programming interface that is accessible to other components in the SensePlace2 UI. This programming interface is used to populate components with data upon query completion and to clean/reset the components between queries, but more importantly, it is used to coordinate user actions across multiple components. In the map interaction example used above, the analyst might be interested in skimming the contents of the tweets that talk about the location they picked on the map. In order to support this, the map component retrieves the list of tweet IDs related to the location of interest, then calls the method:

```
tweetList.highlightTweetsByTweetIds
```

This method is provided by the programming interface of the tweet list component. All that is necessary to make this happen is a bit of code that tells the map component what method from the tweet list programming interface to use. If a two-way coordination is required (e.g., the map component should highlight the locations mentioned in the tweets selected in the tweet list), a bit of code is added to the tweet list component that calls the method:

```
mapComponent.highlightFeatureByTweetId
```

This method is provided by the programming interface of the map component. If programming interfaces of both components are well-defined, this sort of coordination is simple to support.

Writing “glue” code in the fashion described above to coordinate components is a straightforward task when the number of components is small. It does, however, become a serious challenge as the number of components grows. For two components, two bits of glue code are necessary. For three components, six bits are required, for four – 12. In general, it will take $n \times (n-1)$ bits of glue code to link n components, which makes this type of a manual approach to component coordination prohibitively expensive – the latest version of SensePlace2 already has nine components, which would call for 72 bits of glue code that would need to be written and maintained. We have instead opted to resolve this problem through the development of a dedicated *coordinator component* as part of the SensePlace2 interface.

4.2 Coordinator Component

The main logic behind the coordinator component is as follows. For each of the components in the SensePlace2 UI:

1. Determine which user actions this component triggers.

   In the example above, whenever the user clicks a point symbol on a map, the map component calls a “highlight” method on the tweet list; it can be said that the map component *triggers* a “highlight” action inside the tweet list component.

2. Determine which user actions this component listens to.

   In the example above, the only reason the map component can trigger a “highlight” action inside the tweet list is because the tweet list exposed a corresponding method in its programming interface; it can be said that tweet list component is *listening* for “highlight” actions.

3. For each type of user action (e.g., “highlight”, “select”, etc.) bind action triggers to action listeners.

   In the example above, the trigger for “highlight” action inside map component will be bound to the “highlight” listener inside the tweet list component.

The three steps outlined above can be performed in a number of ways. Software components are not “animate” in that they cannot be “asked” for their triggers and listeners directly, and some sort of software mechanism is required to make this kind of introspection possible. The SensePlace2 coordinator uses the metadata approach, where the triggers and listeners of each individual component are described as metadata in a dedicated property of the same component. The name of such a property (`coordinationMetadata`) as well as its inner structure is copied verbatim between components, which makes automatic coordination possible. Shown below is an example of the coordination metadata property from the map component:

```
mapComponent.coordinationMetadata = {

     // Events this component would like to listen to
    listeners: {
        "highlight": mapComponent.onHighlight
    },



    // Events this component triggers and list of their subscribers
    triggers: {

        "select": {
            // List of callbacks is populated by Coordinator.
        },

        "highlight": {
            // List of callbacks is populated by Coordinator.
        }
    }

};
```

Data in the example above will be interpreted by the coordinator component as follows:

1. mapComponent is listening to the “highlight” action and requests its `onHighlight` method to be triggered whenever “highlight” action occurs in any other UI component.

2. mapComponent triggers “select” and “highlight” actions internally and would like to trigger them across the rest of the UI.

Once the coordinator retrieves such data from the rest of the SensePlace2 components, it will make necessary modifications to the “triggers” field of their coordination metadata property, which would, in effect, bind action triggers to action listeners across the entire UI.

4.3 Advantages of the SensePlace2 Coordination Mechanism

The Coordination mechanism described above has a number of useful properties.

First, the coordinator component is only used at the startup of the entire application and then it "steps aside" to let components communicate to each other directly. Thus, the coordinator component does not add performance bottlenecks to the system.

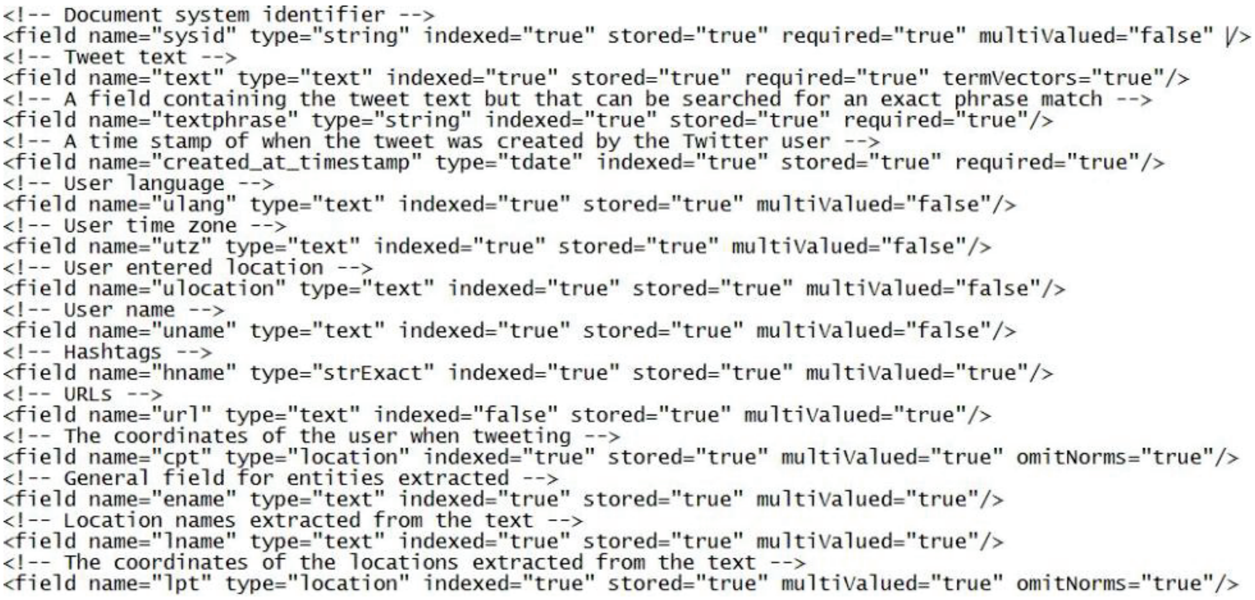
Second, components are free to do as little or as much coordination as desired. For example, one component can trigger a "select" action and listen to nothing, while another would trigger and listen to both "select" and "highlight" actions.

Third, it is easy to add new action types to the system without breaking existing components. Each individual component need only know and care about the types of actions they choose to support, and are completely unaware of what is happening elsewhere. Currently, only "select" and "highlight" actions are used.

Fourth, the coordination mechanism makes sure that all of the SensePlace2 components are made aware of user actions as they happen, but it is up to each individual component to decide what to do with this information. This means that a number of useful and atypical components can be devised. One of these atypical components currently used in the SensePlace2 is a cross-window event gate that provides for transparent communication between multiple browser windows. A separate coordination mechanism is set up in each of the browser windows along with an event-gate component. Event gates communicate with each other, effectively relaying all actions they "snooped" in their respective window to all other windows. Other options include the potential to create a "black box" component that stores all of the user's interactions for later "replay" of the analytical session as well as a cross-computer event gate that can be used to communicate actions across multiple computers during collaborative analytical sessions.

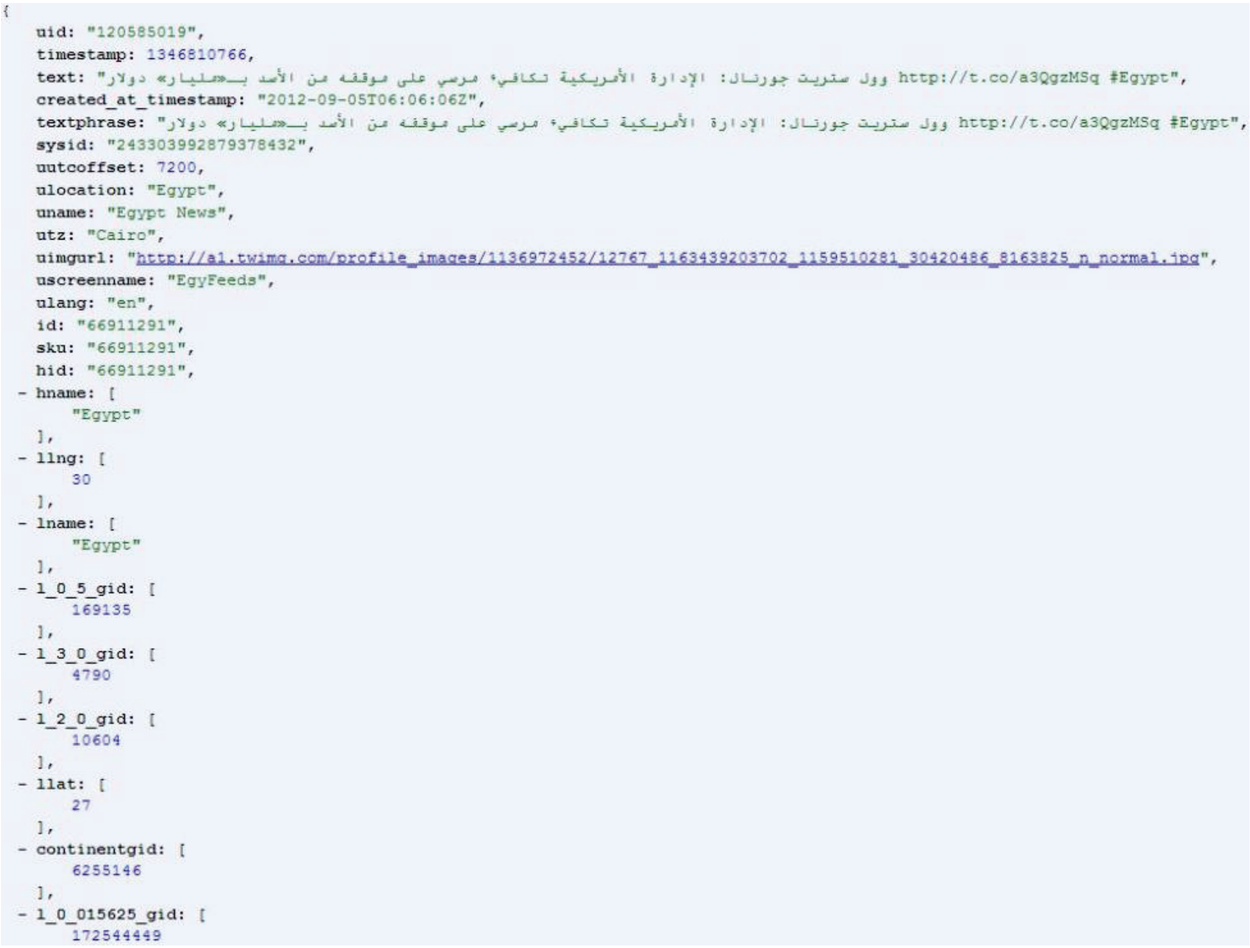
4.4 SOLR and Faceted Search

Advances to the SensePlace2 client-side component coordination described above are supported by advances to the server-side data collection, processing, indexing, and query support mechanisms. The major development in server-side support in SensePlace2 for the Task 1 place-focused analysis is integration of Apache Solr (http://lucene.apache.org/solr/) to index tweet data collected. Solr is an open source application developed by the Apache Software Foundation that has been used in large commercial production environments (e.g., Netflix, Zappos). It provides an efficient means to search text-based information. The Solr approach relies on a schema of fields within a document. As implemented for SensePlace2, each document consists of tweet text and its corresponding metadata. A portion of this schema and an example tweet document returned from a search are shown below in Figure 9 and Figure 10. Within Figure 9, descriptive comments have been added with "<!-- -->" surrounding. Each Solr tweet document will have these fields along with other information fields.



```
<!-- Document system identifier -->
<field name="sysid" type="string" indexed="true" stored="true" required="true" multiValued="false" |/>
<!-- Tweet text -->
<field name="text" type="text" indexed="true" stored="true" required="true" termVectors="true"/>
<!-- A field containing the tweet text but that can be searched for an exact phrase match -->
<field name="textphrase" type="string" indexed="true" stored="true" required="true"/>
<!-- A time stamp of when the tweet was created by the Twitter user -->
<field name="created_at_timestamp" type="tdate" indexed="true" stored="true" required="true"/>
<!-- User language -->
<field name="ulang" type="text" indexed="true" stored="true" multiValued="false"/>
<!-- User time zone -->
<field name="utz" type="text" indexed="true" stored="true" multiValued="false"/>
<!-- User entered location -->
<field name="ulocation" type="text" indexed="true" stored="true" multiValued="false"/>
<!-- User name -->
<field name="uname" type="text" indexed="true" stored="true" multiValued="false"/>
<!-- Hashtags -->
<field name="hname" type="strExact" indexed="true" stored="true" multiValued="true"/>
<!-- URLs -->
<field name="url" type="text" indexed="false" stored="true" multiValued="true"/>
<!-- The coordinates of the user when tweeting -->
<field name="cpt" type="location" indexed="true" stored="true" multiValued="true" omitNorms="true"/>
<!-- General field for entities extracted -->
<field name="ename" type="text" indexed="true" stored="true" multiValued="true"/>
<!-- Location names extracted from the text -->
<field name="lname" type="text" indexed="true" stored="true" multiValued="true"/>
<!-- The coordinates of the locations extracted from the text -->
<field name="lpt" type="location" indexed="true" stored="true" multiValued="true" omitNorms="true"/>
```

Figure 8: SensePlace2 partial Solr schema (with comments)



```
{
   uid: "120585019",
   timestamp: 1346810766,
   text: "وول ستريت جورنال: الإدارة الأمريكية تكافيء مرسي على موقفه من الأسد بـ«مليار» دولار http://t.co/a3QgzMSq #Egypt",
   created_at_timestamp: "2012-09-05T06:06:06Z",
   textphrase: "وول ستريت جورنال: الإدارة الأمريكية تكافيء مرسي على موقفه من الأسد بـ«مليار» دولار http://t.co/a3QgzMSq #Egypt",
   sysid: "243303992879378432",
   uutcoffset: 7200,
   ulocation: "Egypt",
   uname: "Egypt News",
   utz: "Cairo",
   uimgurl: "http://a1.twimg.com/profile_images/1136972452/12767_1163439203702_1159510281_30420486_8163825_n_normal.jpg",
   uscreenname: "EgyFeeds",
   ulang: "en",
   id: "66911291",
   sku: "66911291",
   hid: "66911291",
 - hname: [
      "Egypt"
   ],
 - llng: [
      30
   ],
 - lname: [
      "Egypt"
   ],
 - l_0_5_gid: [
      169135
   ],
 - l_3_0_gid: [
      4790
   ],
 - l_2_0_gid: [
      10604
   ],
 - llat: [
      27
   ],
 - continentgid: [
      6255146
   ],
 - l_0_015625_gid: [
      172544449
```

Figure 9: SensePlace2 partial returned document from a Solr search

In addition to a simple text-based search, one of the more important features of Solr is *faceting* (faceting partitions data into sets and subsets for quick drill-down and retrieval, e.g., Tang, 2007; Yee et al., 2003). In SensePlace2, data are currently partitioned into the following high level facets (in bold) and subfacets (in plain text).

Free text
User entered text
From location
Yes/No
Time
User selected time span
Overview location
User selected grid cells
Hierarchy locations
User selected countries
Location Distance
Distance away from the user selected location

As an example application of a facet-based query, the user might start with a query for "protest", which isolates the subset of all tweets containing that term. Next, the user might select a time-frame that drills-down to tweets about protest and within that time-frame. Lastly, the user can subset again geographically based on the subset of tweets with protest within the time-frame that contains *about* places in a particular country. As the user applies facets, it is important to provide feedback on the frequency of hits in the data set contained in the result. Within SensePlace2, Solr provides on-the-fly aggregate counts of all the data matching the search parameters. In the SensePlace2 maps, locations mentioned in the most relevant 1000 tweets are depicted as points on the map scaled to represent frequency tertiles. To get the overview of the rest of the locations, a gridded cartographic heatmap is also displayed. The grid cell counts for each heatmap cell are also generated on-the-fly by Solr.

Beyond using Solr to support faceted query, SensePlace2 leverages Solr's support for boosting search results as a component in generating and sorting the 1000 most relevant tweets matching any user query. In addition, Solr is capable of grouping documents similar to each other. Because many tweets are very similar to one another, SensePlace2 uses this capability to return a representative example of groups of similar tweets, thus avoiding redundancy in what the user examines by hand. Finally, SensePlace2 makes use of Solr geographic sorting capabilities (e.g., to sort tweets by their *about* or *from* distance from a user selected location).

5 Conclusions and Future Work

This report presents the Task 1 outcomes from our larger research directed to supporting information foraging and sensemaking with open media; Task 1 focuses on place-based analysis. The report contains three components.

First, we outline findings about the types of place-related information that is contained explicitly or can be derived from Twitter posts (obtained using the Twitter streaming API that provides selected posts plus metadata for each). Important insights of this aspect of the work include: (a) identification of substantial thematic variability in the propensity in both tweets with geolocation (for which users have opted in to providing their location) and tweets containing place references in the text, (b) determination that (for our U.S. sample) most (about 91%) of geolocated tweets contain coordinates with about 9% providing only named places without coordinates. On average about 1.5% of tweets included geolocation, but some crisis events (earthquake and flood) were found to have substantially higher percentages, which suggests that some individuals consciously opt-in to include geolocation in times of crisis. In addition, 10% of tweets in our sample included place references in the text and that percentage varied substantially across topics (over 50% of tweets for the search term "protest" included place references).

Second, we present visual-analytics methods that were implemented (or extended) in support of place-based analysis that draws on a large and continually growing (200 million plus) repository of Twitter tweets. Most of the methods presented have been implemented as interlinked components within SensePlace2 (SensePlace2), a web-based application designed to support information foraging and sensemaking from unstructured text (SensePlace2 was initially developed to support situational awareness for crisis management with funding from the Department of Homeland Security and continued funding from that source has emphasized development of a stand-alone API for recognizing, disambiguating, and geolocating references to place in short text "documents", including tweets). Work reported here extended prior tools for filtering data on the basis of a formal hierarchy of geographic names and for highlighting frequently referenced place names in place tag clouds. Emphasis in research presented here has been on developing strategies to uncover and represent links among places and links of other features to place. Map-based co-referencing allows an analyst to quickly identify the places in the world that are referenced together with any place of interest (by pointing to the place on the SensePlace2 map). A Co-occurrence Matrix has also been added as a more general tool for depicting co-occurrence. It supports not only depiction of frequency of place-place co-mentions but also place-named region, place-person, place-day of the week, and place-hashtag analysis. In addition to visual methods implemented within SensePlace2, we also introduce ideas for future work to support movement analysis based upon geolocated tweets. That preliminary work has produced an initial application that can monitor movements reflected in tweets in real time and play back past movements in compressed time.

To enable the functionality above, many of the data query capabilities that have been implemented in SensePlace2 as database functionality have been migrated to leverage Solr-Lucene indexing and query capabilities. This has made it possible to scale functionality to larger volumes of data and to more easily implement drill-down to focused subsets of information. While that change in system architecture has been important for system performance, the more innovative aspect of our system engineering work is development and implementation of a strategy to support software component coordination in a browser using JavaScript. This advance supports not only flexible brushing and linking among views in a single browser (e.g., highlighting on the map propagates to the Tweet list and place cloud), but also cross-browser coordination. As a result, different visualization components that launch in different browser windows can support dynamic linking among the views. The architecture of the approach will enable remote collaboration among distributed users whose browser windows can interact.

The scale of data produced by Twitter and other open sources of unstructured text presents a range of challenges for visual analytics tools that can enable flexible place-based (or organization, person, social network -based) information foraging, sensemaking, and situational awareness. Unlike many data sources, Twitter produces data that are dense in both space and time. The research reported here has generated some advances in both knowledge of the data and methods to leverage those data. Multiple open questions and challenges remain. Among these, we highlight 3 that present particular opportunities to advance our ability to leverage geographic information from big, open, unstructured text sources.

First, current methods to recognize, disambiguate, and geolocate place-relevant information in text (particularly microblog text) produce many errors. While progress is being made on improving the accuracy of these methods, improved accuracy alone is not sufficient due to the scale of data involved. Improving both the accuracy and the speed of methods is necessary to support real analysis. Promising strategies to address this include enhancing approaches to utilizing spatial, temporal, social, and event context; leveraging user input that can provide real-time clues to disambiguation methods; and leveraging advances in cloud computing to distribute processes that must be applied repeatedly.

Second, there is a need for new visual interfaces to explore the place-related facets of the information we are now able to collect and evaluate. One specific goal is to develop a temporal view of place references in order to be able to identify which places are trending. At present, we can examine which places are relevant to a keyword search, but we are not able to look at temporal patterns in geographic references that might reveal patterns in certain places. Because we are able to leverage the hierarchies used in Geonames, it should be possible to design and develop a graph view that shows countries, continents, cities, and other geographic references and their relative frequencies over time as well as any deviations from “normal” for particular places. This could be filtered by co-occurrence with other keywords, or we could simply explore the emergent trends in geographic references alone.

Third, analysis is often a complex process that is carried out over extended periods of time. Methods are needed to support analytical threads as events happen, relationships are identified, and patterns in space and time are recognized. Supporting analysis over time requires methods for information and knowledge capture and management. Here, we have made a step toward supporting the analytical process over time through the component coordination mechanism introduced with its ability to add a component that listens to all other components to capture the process of analysis for later replay and analysis. A goal for future research is to integrate this capability with knowledge management methods to support an analytical process that may be distributed among devices and users.

Acknowledgement:

We extend our appreciation to GeoVISTA Ph.D. student Eun-Kyeong Kim who was responsible for analyzing the Geolocation Stream data set. Tables 1a-c and our discussion related to results from analyzing those data were possible due to her comprehensive analysis of the data set.

6 References

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M. and Pölitz, C. 2010: Discovering Bits of Place Histories from People's Activity Traces. *IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2010),* Salt Lake City, Utah, USA, 59-66.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D.S. and Ertl, T. 2012: Spatiotemporal Social Media Analytics for Abnormal Event Detection using Seasonal-Trend Decomposition. *IEEE VAST,* Seattle: IEEE.

Chandra, S., Khan, L. and Muhaya, F. 2011: Estimating twitter user location using social interactions--a content based approach. *IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing (PASSAT/SocialCom 2011),* Boston, Massachusetts, USA: IEEE, 838-843.

Cheng, Z., Caverlee, J. and Lee, K. 2010: You are where you tweet: a content-based approach to geo-locating twitter users. *Proceedings of the 19th ACM international conference on Information and knowledge management,* Toronto, Ontario, Canada: ACM, 759-768.

Crooks, A., Croitoru, A., Stefanidis, A. and Radzikowski, J. 2012: #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17(1): 124-147.

Davis, C.A., Pappa, G.L., de Oliveira, D.R.R. and de L. Arcanjo, F. 2011: Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS* 15, 735-751.

Diakopoulos, N., Naaman, M. and Kivran-Swaine, F. 2010: Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry. *IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2010),* Salt Lake City, Utah, USA, 115-122.

Dörk, M., Carpendale, S., Collins, C. and Williamson, C. 2008: VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *Information Visualization 2008,* Columbus, OH, 1205-1212.

Dörk, M., Gruen, D., Williamson, C. and Carpendale, S. 2010: A Visual Backchannel for Large-Scale Events. *IEEE Transaction on Visualization & Computer Graphics* 16, 1129-1138

Field, K. and O'Brien, J. 2010: Cartoblography: Experiments in Using and Organising the Spatial Context of Micro-blogging. *Transactions in GIS* 14, 5-23.

Gelernter, J. and Balaji, S. 2013: An algorithm for local geoparsing of microtext. *GeoInformatica*, 1-33.

Gelernter, J. and Mushegian, N. 2011: Geo-parsing Messages from Microtext. *Transactions in GIS* 15, 753-773.

Gonzalez, R., Figueroa, G. and Chen, Y.-S. 2012: TweoLocator: a non-intrusive geographical locator system for Twitter. *Proceedings of the 5th International Workshop on Location-Based Social Networks,* Redondo Beach, California: ACM, 24-31.

Guy, M., Earle, P., Ostrum, C., Gruchalla, K. and Horvath, S. 2010: Integration and Dissemination of Citizen Reported and Seismically Derived Earthquake Information via Social Network Technologies. In Cohen, P., Adams, N. and Berthold, M., editors, *Advances in Intelligent Data Analysis IX*: Springer Berlin Heidelberg, 42-53.

Hecht, B., Hong, L., Suh, B. and Chi, H. 2011: Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles. *CHI 2011,* Vancouver, BC, Canada.

Ikawa, Y., Vukovic, M., Rogstadius, J. and Murakami, A. 2013: Location-based insights from the social web. *Proceedings of the 22nd international conference on World Wide Web companion*: International World Wide Web Conferences Steering Committee, 1013-1016.

Jie, Y., Lampert, A., Cameron, M., Robinson, B. and Power, R. 2012: Using Social Media to Enhance Emergency Situation Awareness. *Intelligent Systems, IEEE* 27, 52-59.

Jones, C.B. and Purves, R.S. 2008: Geographical information retrieval. *International Journal of Geographical Information Science* 22, 219 - 228.

Kent, J.D. and Capello, H.T. 2013: Spatial patterns and demographic indicators of effective social media content during theHorsethief Canyon fire of 2012. *Cartography and Geographic Information Science* 40, 78-89.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A. and Shook, E. 2013: Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday, [S.l.], apr. 2013. ISSN 13960466. Available at: <http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654>. Date accessed: 28 Aug. 2013. doi:10.5210/fm.v18i5.4366* 18.

Li, W., Serdyukov, P., de Vries, A.P., Eickhoff, C. and Larson, M. 2011: The Where in the Tweet. *CIKM'11,* Glasgow, Scotland, UK.

Lieberman, M.D. and Samet, H. 2012: Adaptive context features for toponym resolution in streaming news. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*: ACM, 731-740.

Lingad, J., Karimi, S. and Yin, J. 2013: Location extraction from disaster-related microblogs. *Proceedings of the 22nd International Conference on World Wide Web Companion*, Rio de Janeiro, Brazil, May 13-17, 2013, 1017-1020.

Liu, X., Wei, F., Zhang, S. and Zhou, M. 2013: Named entity recognition for tweets. *ACM Trans. Intell. Syst. Technol.* 4, 1-15.

MacEachren, A.M., Jaiswal, A., Robinson, A.C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X. and Blanford, J. 2011a: SensePlace2: GeoTwitter Analytics Support for Situational Awareness. In Miksch, S. and Ward, M., editors, *IEEE Conference on Visual Analytics Science and Technology,* Providence, RI: IEEE, 181 - 190.

MacEachren, A.M., Robinson, A.C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J. and Mitra, P. 2011b: Geo-Twitter Analytics: Applications in Crisis Management. *25th International Cartographic Conference,* Paris, France.

Mahmud, J., Nichols, J. and Drews, C. 2012: Where is this tweet from? inferring home locations of twitter users. *Proc AAAI ICWSM* 12.

Mendoza, M., Poblete, B. and Castillo, C. 2010: Twitter Under Crisis: Can we trust what we RT? , *1st Workshop on Social Media Analytics (SOMA '10),* Washington, DC, USA.

Mirkovic, M., Culibrk, D. and Crnojevic, V. 2012: Mining Geo-Referenced Community-Contributed Multimedia Data. In Abraham, A., editor, *Computational Social Networks*: Springer London, 81-102.

Morstatter, F., Kumar, S., Liu, H. and Maciejewski, R. 2013: Understanding Twitter Data with TweetXplorer. *KDD '13,* Chicago, Illinois.

Oh, O., Agrawal, M. and Rao, H. 2010: Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. *Information Systems Frontiers*, 1-11.

Sixto, J., Pena, O., Klein, B. and López-de-Ipina, D. 2013: Enable tweet-geolocation and don't drive ERTs crazy! Improving situational awareness using Twitter. *Proceedings of SMERST 2013: Social Media and Semantic Technologies in Emergency Response,* University of Warwick, Coventry UK, 27-31.

Sreenivasan, N., Lee, C. and Goh, D. 2011: Tweet me home: exploring information use on twitter in crisis situations. *Online Communities and Social Computing*, 120-129.

Tang, M.-C. 2007: Browsing and searching in a faceted information space: A naturalistic study of PubMed users' interaction with a display tool. *Journal of the American Society for Information Science and Technology* 58, 1998-2006.

Thom, D., Bosch, H., Koch, S., Worner, M. and Ertl, T. 2012: Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, 41-48.

Tomaszewski, B. and MacEachren, A.M. 2010: Geo-Historical Context Support for Information Foraging and Sensemaking: Conceptual Model, Implementation, and Assessment. *IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2010),* Salt Lake City, Utah, USA, 139-146.

Tomaszewski, B. and MacEachren, A.M. 2012: Geovisual Analytics to Support Crisis Management: Information Foraging for Geo-Historical Context. *Information Visualization {invited extension of paper originally published in Proceedings of IEEE VAST 2010}* 11, 339-359.

Tsou, M.-H., Yang, J.-A., Lusher, D., Han, S., Spitzberg, B., Gawron, J.M., Gupta, D. and An, L. 2013: Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, 40(4): 1-12.

Tsukayama, H. 2013: Twitter turns 7: Users send over 400 million tweets per day. *The Washington Post,* Washington, D.C.: The Washington Post.

Vieweg, S., Hughes, A., Starbird, K. and Palen, L. 2010: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. Proc. of the 28th Inter. Conference on Human Factors in Computing Systems: ACM, 1079-1088.

Yee, K.P., Swearingen, K., Li, K. and Hearst, M. 2003: Faceted metadata for image search and browsing. Proceedings of the SIGCHI conference on Human factors in computing systems, Ft. Lauderdale, FL, USA, April 5-10, 2003 , 401-408.

Zheng, Y.-T., Li, Y., Zha, Z.-J. and Chua, T.-S. 2011: Mining Travel Patterns from GPS-Tagged Photos. In Lee, K.-T., Tsai, W.-H., Liao, H.-Y., Chen, T., Hsieh, J.-W. and Tseng, C.-C., editors, Advances in Multimedia Modeling: Springer Berlin Heidelberg, 262-272.